Construyendo un lexicón bilingüe anglo-español para nombres propios utilizando la Web

Autor: Ivan Lyubomirov Pavlov

Universidad: The University of Sheffield

Tutor: Dr. Robert Gaizauskas

Cotutor: Jorge Blasco

Coordinador Erasmus: Jose Maria Fuentes

Fecha de lectura: 18/01/2010

Calificación: 80/100

Contents

1 Resumen

El uso de diccionarios electrónicos es una tarea común en sistemas de búsqueda translingüe (del inglés Cross Language Information Retrieval CLIR) y traducción automática (TA). Existen muchos diccionarios electrónicos, pero estos no incluyen nombres de entidades. Además cada día aparecen nuevos nombres. Dado que los nombres son una parte esencial en las búsquedas translingües y en el lenguaje en general es necesario buscar una forma de traducirlos de un idioma a otro. El objetivo de este proyecto es construir un lexicón de nombres de entidades obteniendo los datos en textos traducidos publicados en la Web. En el caso perfecto el lexicón se actualizará constantemente utilizando los recursos en continuo crecimiento de la Web.

Este documento es un resumen en español del proyecto original en inglés que contiene un estudio de campo que ha llevado al diseño e implementación de un sistema que construye un lexicón de nombres propios, utilizando artículos de noticias escritos en inglés y español y descargados de la Web. El lexicón contiene más de 10000 nombres obtenidos en más de 27000 parejas de artículos. El sistema tiene la capacidad de actualizar el contenido del lexicón utilizando los nuevos artículos que se publican cada día.

2 Introducción

2.1 Antecedentes

Buscar en la Web se ha convertido en una parte del día a día. Pero la capacidad de búsqueda de la gente está limitada por factores como el idioma. En un escenario perfecto la frase de búsqueda se escribiría en un idioma y los resultados serían documentos relevantes en cualquier idioma. La TA es tan avanzada hoy en día, que no hace falta saber el idioma original, ya que la máquina puede traducirlo. Para poder hacer realidad este escenario hay mucha investigación en el campo de la búsqueda translingüe y la TA. Estas dos técnicas se apoyan en diccionarios bilingües para sus traducciones. Sin embargo estos diccionarios no contienen nombres propios como nombres de personas y organizaciones. No obstante estos nombres son frecuentemente incluidos en las búsquedas o incluso son toda la frase de búsqueda. Normalmente estos nombres coinciden en idiomas diferentes, especialmente cuando estos idiomas utilizan el mismo alfabeto como el español y el inglés, pero esto no siempre es así. Los nombres de sitios a veces son diferentes como por ejemplo Londres es London en inglés. Lo mismo pasa con los nombres propios, el nombre original del príncipe Carlos es Charles. Otro ejemplo son los nombres de películas, que se podrían considerar como texto, pero incluso los traductores humanos pueden tener dificultades en su traducción, ya que pocas veces se traducen literalmente. Así que para traducirlos hay que saber cómo se llama la película en ambos idiomas. Además constantemente se introducen nuevos nombres haciendo muy difícil incluirlos en un diccionario de nombres.

2.2 Objetivos

El principal objetivo del proyecto es construir un lexicón de nombres para su traducción entre inglés y español. El lexicón tiene que ser construido automáticamente utilizando textos bilingües obtenidos en la Web. Uno de los objetivos segundarios es que el lexicón se actualice continuamente con nuevos nombres que aparecen cada día en la red. La restricción principal del proyecto es el tiempo, ya que este se desarrolla en un único cuatrimestre.

Para cumplir los objetivos hay que diseñar un sistema que localiza páginas traducidas en español e inglés construyendo un corpus paralelo. Luego el sistema tiene que localizar los nombres propios en los dos textos y emparejarlos para obtener el nombre en español y su traducción en inglés.

Hay tres partes claras en el proyecto y hay varias alternativas para resolver cada una de las tres. La primera parte es la construcción del corpus. Algunas de las diferentes alternativas son la utilización de algún corpus existente o la implementación de una araña web (crawler) capaz de localizar las páginas traducidas. Una alternativa a los textos traducidos es un corpus comparable que consiste de documentos que no son traducciones pero tienen el mismo tema. Por ejemplo las agencias de noticias suben en la Web todos los días documentos con información similar, escritos en idiomas diferentes.

La segunda parte del proyecto consiste en reconocer los nombres propios en el texto en ambos idiomas. Las alternativas para hacerlo consisten de simples heurísticas a técnicas de aprendizaje automático. También existen diferentes programas gratuitos y de código abierto que podrían ayudar en la tarea.

La última parte es emparejar los nombres en un idioma con los nombres del otro. Habiendo localizado los nombres, hay que alinear los textos para emparejarlos. Algunas heurísticas o modelos estadísticos pueden ser utilizados para cumplimentar esta tarea final.

3 Investigación

En esta sección se discuten otros proyectos sobre el mismo tema o temas similares. El objetivo es investigar diferentes alternativas para el cumplimiento de los objetivos del proyecto. La sección está dividida en tres partes, una para cada una de las partes del proyecto.

3.1 Construcción del corpus bilingüe

El primer paso del proyecto es obtener muchos textos alineados en inglés y español, que tengan alto contenido de nombres propios. Hay mucha literatura y diferentes trabajos en este campo.

Sin embargo uno de los proyectos más importantes en este campo es STRAND (Structural Translation Recognition for Acquiring Natural Data). Este sistema obtiene de la Web de forma automática parejas de textos en inglés y español. Hay dos versiones del sistema. Una de las estrategias de la primera versión era utilizar el famoso buscador AltaVista para buscar páginas en francés que contienen enlaces llamados “English” o “anglias” y páginas en inglés que contienen enlaces llamados “French” o “francias”. Suponiendo que estos enlaces llevan a la traducción de la página se obtiene una colección de parejas candidatas de enlaces en ambos idiomas.

Después de obtener las parejas candidatas se llevan a cabo diferentes operaciones estadísticas para identificar si las parejas son buenas (traducciones de inglés al francés o viceversa) o no.

La siguiente versión de STRAND utiliza un sitio web llamado Internet Archive (http://www.archive.org/). El “Archivo de Internet” es una organización sin beneficios que intenta almacenar copias de toda la Web pública en diferentes momentos del tiempo.

En vez de utilizar AltaVista para la generación de parejas candidatas se utiliza el Archivo de Internet. Para generar las parejas se buscan ahí distintas URL-es que sólo se diferencian en partes que pueden designar el idioma de la página.

Otro sistema relevante es BITS. Este sistema utiliza un simple algoritmo para generar las parejas candidatas haciendo una query a un servidor DNS en busca de sitios en el dominio del menor de los dos idiomas. La lista de sitios es explorada para identificar si son bilingües. Si lo son, se descargan todas las páginas del dominio y luego se filtran las parejas, comparando el tamaño de las páginas y el número de enlaces y haciendo comparación del contenido.

Otra forma completamente diferente de obtener un corpus bilingüe se puede encontrar en el siguiente proyecto, que se apoya en agencias de noticias internacionales que publican sus noticias en varios idiomas, traduciendo el articulo original (normalmente en inglés) a otros idiomas de interés. Lo que hace el sistema muy fácil de implementar es la utilización de RSS para publicar las noticias. El sistema descarga las noticias vía RSS y luego descarga la versión original cuyo enlace está contenido dentro de la noticia. El éxito de este método depende de la búsqueda manual de sitios web que publican contenidos en los dos idiomas de interés.

En otros artículos se propone la utilización de textos comparables en vez de paralelos para fines similares al fin de este proyecto. Los textos comparables son más fáciles de obtener y más abundantes ya que hay muchos artículos sobre el mismo tema que no son necesariamente traducciones. En estos trabajos también se utilizan noticias para la construcción del corpus.

Otra opción para el proyecto es la utilización de algún corpus existente. Se encuentran varios corpus multilingües que contienen textos traducidos en inglés y español, que se pueden utilizar libremente. Estos son Europal – que contiene textos del parlamento europeo, JRC-Acquis – toda la ley de la UE, OPUS – varios corpus entre ellos uno de documentos médicos y otro de subtítulos de películas.

3.2 Reconocimiento de nombres propios

La segunda parte del proyecto, después de obtener el corpus bilingüe, es extraer todos los nombres del texto. Esta tarea se conoce por el nombre **reconocimiento de nombres de entidades RNE**, *Named Entity Recognition NER.* La investigación en este campo comienza en el año 1991 cuando se empezó a creer interesante localizar nombres de personas, sitios y organizaciones de forma automática. En el principio se utilizaban reglas básicas pero luego se introducen las técnicas de aprendizaje automático. Un estudio de los trabajos desarrollados en este área se pueden encontrar en (1). Este trabajo da una idea clara de la dificultad de la tarea de RNE. Esta parte del proyecto es muy importante, porque los malos resultados en esta fase del proyecto afectarían severamente el resultado final del sistema. Por eso, dado el poco tiempo para el desarrollo del sistema, en vez de investigar las técnicas de RNE se investigan sistemas ya desarrollados que llevan a cabo esta tarea y que son gratuitos.

Algunos de estos sistemas son Balie, FreeLing, JBL Named Entity Tagger y AlchemyAPI. Cada uno de ellos tiene la capacidad de reconocer nombres de entidades tanto en inglés como en español.

3.3 Alineamiento de nombres

La parte final del proyecto recibe como entrada los nombres de los textos en ingles y los de los textos en español y tiene que producir parejas de nombres que se colocarán en el lexicón.

Uno de los trabajos más importantes sobre el alineamiento de palabras para traducción automática es el proyecto de IBM. Aunque es bastante antiguo, este trabajo es referenciado en muchos de los proyectos en este área y algunos de los métodos se usan todavía. Una descripción detallada de los modelos IBM de 1 al 5 se puede encontrar en (2).

Para este proyecto se considera interesante el primer modelo. IBM 1 es un modelo estadístico que calcula las probabilidades de las diferentes traducciones para una frase dada basándose en las probabilidades de traducción de cada una de las palabras. Lo interesante es como el algoritmo de maximización de la expectación ME, *expectation maximization EM,* se aplica al modelo. Este algoritmo recibe como entrada una lista de frases traducidas y calcula la probabilidad de que una palabra extranjera *f* sea traducida a una palabra española *e* para cada una de las palabras de entrada.

El modelo IBM 1 es un buen modelo para empezar, pero tiene muchas deficiencias. Por ejemplo no permite traducir una palabra como varias o viceversa. Además no da información sobre el orden de las palabras. Para eso están los modelos del 2 al 5. IBM 2 incluye un modelo explicito para la alineación de las palabras. IBM 3 permite que una palabra sea traducida como varias o ninguna. IBM 4 añade alineamiento relativo e IBM 5 arregla deficiencias. Todos estos modelos son construidos sobre los modelos anteriores preservando los principios generales.

4 Análisis de requisitos

El requisito principal del proyecto es construir un lexicón para traducir nombres entre inglés y español, utilizando la Web. Para que el lexicón sea útil, tiene que cubrir diversos dominios y tiene que ser actualizado regularmente. El análisis se divide en tres partes.

4.1 Construcción del corpus bilingüe

La primera parte es obtener un corpus Anglo-Español para aprender de ahí las traducciones de los nombres. Para que el lexicón cubra diferentes dominios los textos del corpus tienen que ser diversos en contenido. Otro requisito es que hay que incrementar el corpus regularmente con nuevos textos.

Después de considerar las diferentes alternativas para la construcción del corpus, se elige utilizar artículos noticiaros de agencias que publican la misma historia en varios idiomas, aunque los artículos no san traducidos. Los artículos de noticias son el ejemplo perfecto de nuevos textos publicados todos los días y con temas diversos. Normalmente la misma noticia es relatada por

diferentes periodistas en idiomas distintos pero incluyendo los mismos hechos y por tanto incluyendo los mismos nombres de sitios, organizaciones o personas. Esto hace la construcción de un corpus comparable muy útil para el cumplimiento de los objetivos del proyecto.

Para la construcción del corpus se utiliza el sitio web de la agencia Euronews. Este sitio contiene un archivo de noticias que data desde el año 2004. Para cada día contenido en el archivo existe una página con URL parecido a este http://www.euronews.net/2004/10/07/ que contiene enlaces a todas las noticias del día. Para los artículos en español de este mismo día los enlaces están en la página http://es.euronews.net/2004/10/07/. Hay una imagen asociada a cada uno de los artículos y la imagen es la misma tanto en la versión inglesa como en la española. Utilizando esta imagen se consigue emparejar los artículos comparables con precisión de 100%. Por otro lado la estructura HTML de cada página es siempre la misma, permitiendo extraer de forma fácil tanto los enlaces a los artículos como los títulos y el texto excluyendo la información innecesaria como otros enlaces o publicidad.

El archivo de Euronews se actualiza todos los días con las noticias del día anterior permitiendo incrementar el corpus a diario.

4.2 Reconocimiento de nombres de entidades

Esta parte del proyecto lleva su único requisito implícito en el nombre. Ya se ha visto que no es factible implementar un sistema de RNE. Por esto se va a utilizar uno de los que ya existen. Para poder elegir el mejor sistema es necesario evaluarlos teniendo el corpus bilingüe, pero la evaluación de un sistema de RNE consume mucho tiempo y por esto no se va a llevar a cabo.

El sistema elegido es Alchemy. Las razones para elegirlo son el lenguaje de programación, el sistema operativo y la facilidad de uso. El lenguaje es Java, el sistema es Windows y se proporciona una API muy simple y bien documentada. Además Alchemy es un sistema de doble licencia (libre y comercial) que lleva a pensar que tiene que ser suficientemente bueno para que alguien considere pagar por utilizarlo, en vez de utilizar alguna de las alternativas gratuitas. Para los objetivos del proyecto se va a usar la licencia gratuita, limitando las llamadas a la API a 30.000 por día que deberían ser más que suficientes.

4.3 Alineamiento de nombres

La última y más importante parte del proyecto es la construcción del lexicón alineando los nombres encontrados en la fase previa. Esta fase es similar a un corpus bilingüe alineado a nivel de oración en el que se tiene que hacer alineación de palabras.

La mayoría de la literatura sobre alineamiento de palabras es para corpus paralelos y no comparables como el modelo 1 de IBM. Por eso las palabras de las oraciones en un idioma aparecen traducidas en el otro, algo que no es necesariamente cierto en un corpus comparable.

En el caso de este proyecto los artículos no siempre incluyen los mismos nombres. Aunque una pareja de artículos tenga el mismo contenido, a veces es posible utilizar dos nombres diferentes dando la misma información, por ejemplo un artículo puede utilizar Obama mientras en otro se puede decir el presidente de los EEUU. Además de esto la entrada a esta fase es la salida del RNE que en un buen escenario tiene un coeficiente de recuperación del 80%. El problema es que haciendo RNE en dos idiomas diferentes los 20% de nombres no reconocidos en un idioma pueden ser diferentes que en el otro.

Además del coeficiente de recuperación la precisión también puede ser problemática. Incluso los mejores sistemas de RNE no tienen precisión de 100%. Baja precisión en la fase RNE produce una entrada de baja calidad para la última fase. La entrada contiene muchas palabras que no son nombres de entidades y que pueden ser confundidas por traducciones de nombres reconocidos correctamente.

Todas estas deficiencias hacen muy inapropiado el uso de técnicas de alineación de corpus paralelos y por eso se diseñará un nuevo algoritmo.

La entrada a esta fase son parejas de conjuntos de nombres, un conjunto en inglés y uno en español. Algunos de los nombres en estos conjuntos van a ser traducciones y otros no. El algoritmo tiene que decidir cuáles son las parejas de traducciones.

Está claro que no se puede tomar la decisión con una sola pareja, ya que no hay suficiente información. Sin embargo cuanto más se repita en diferentes artículos la mutua ocurrencia de un NE en español y otro NE en inglés más probable es que esto sea una traducción. Ésta es una simple idea que se va a utilizar para alinear los NE.

Por otro lado una característica muy importante de la entrada es que muchos de los nombres en un idioma van a ser exactamente iguales en el otro, porque los nombres de personas se escriben casi siempre igual en inglés y español. Aun cuando no son exactamente iguales muchos de ellos son muy similares. Por eso el primer criterio para tomar la decisión va a ser un coeficiente de similitud. Para cuantificar la similitud de dos cadenas se utiliza el algoritmo de Jaro Winkler implementado en la librería de código abierto Sam's String Metrics.

4.4 Integración de las partes

El proyecto está dividido en tres partes que se tienen que integrar en un único sistema. Esto significa que la salida de una fase va a ser la entrada a la siguiente. La salida de la última fase es el lexicón. Esto se tiene en consideración para diseñar cada una de las partes.

Por otro lado las partes se tienen que implementar como módulos separados para que se puedan cambiar. Aunque no se cambien en este proyecto como futuro trabajo puede ser interesante cambiar por ejemplo el sistema de RNE con algún otro sin tener que modificar las otras partes del sistema.

La salida de cada una de las fases del sistema tiene que ser persistente. La idea es que si se hace necesario modificar la última fase no exista la necesidad de ejecutar las dos primeras.

5 Diseño, implementación y pruebas

El lenguaje de programación a utilizar es Java. El programador es muy familiar con este lenguaje y los programas de terceros que se utilizan son también escritos en Java.

5.1 Construcción del corpus

La construcción del corpus se divide en dos partes. Primero se generan las URL-es del archivo empezando por http://www.euronews.net/2004/10/07/ y llegando hasta la fecha de ayer y se guardan todos los enlaces a artículos junto con el enlace a la foto del artículo. Esto se hace para la versión en inglés y para la de español. La foto se utiliza para emparejar los enlaces en español con los de inglés. Se guarda una lista en un fichero de texto con las parejas de artículos.

En la segunda parte se descarga el texto de cada uno de los artículos y se almacena en un fichero separado. La estructura HTML es siempre la misma así que es fácil de extraer solo el texto de la noticia.

Los artículos se almacenan en un fichero zip.

Para que el corpus pueda ser incrementado con las nuevas entradas del archivo en cada ejecución se almacena la última fecha explorada para que en la siguiente ejecución se empiece desde la siguiente fecha.

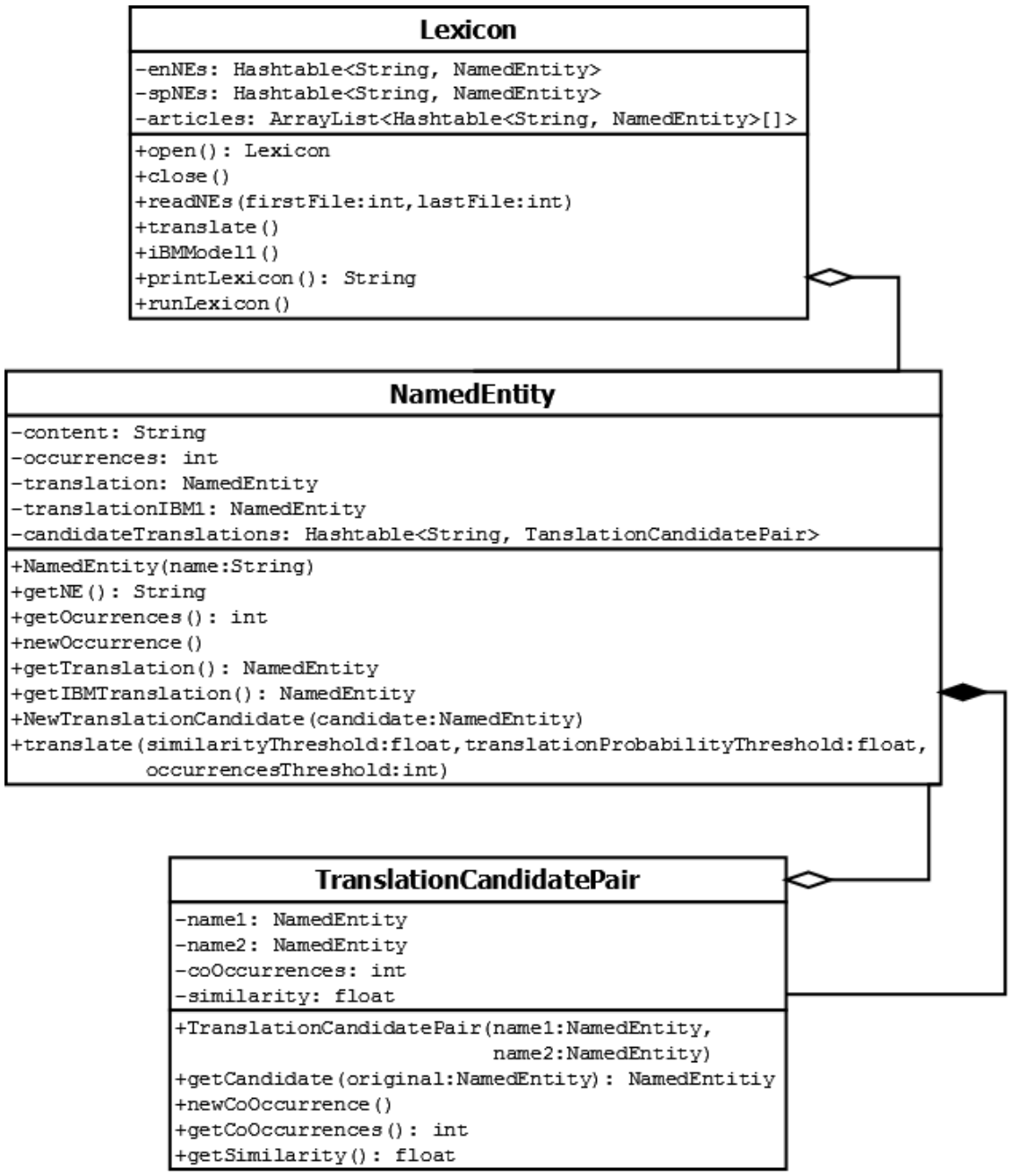
5.2 Reconocimiento de nombres de entidades

La API de Alchemy proporciona una función que recibe una cadena de texto y devuelve una cadena en formato XML con los NE extraídos de la entrada. Esta salida se almacena en archivos incluidos en un fichero zip y luego se procesa para extraer sólo los nombres y se crea un archivo con un nombre por línea para cada uno de los artículos. Estos archivos se almacenan en otro fichero zip.

5.3 Alineamiento de nombres

El fin de esta fase es leer todos los nombres en el sistema sin perder información. Lo que hay que guardar son los nombres en ingles, los nombres en español, el número de veces que aparecen, las traducciones candidatas, el número de coocurrencias de un nombre y todas sus traducciones candidatas y los nombres que aparecen en cada artículo. Una traducción candidata de un nombre en español **e** son todos los nombres en inglés que aparecen en la versión inglesa del artículo español donde aparece **e**.

El diseño consiste en tres clases que se ven en la figura 5.1. La clase más importante es NamedEntity que representa un NE. Esta clase contiene la información sobre el número de veces que aparece el nombre y una Hashtable con todas las traducciones candidatas. Las entradas en la tabla hash son de la clase TranslationCandidatePair. Una pareja de traducciones candidatas consiste en dos referencias a nombres, una en inglés y otra en español, y en número de veces que ocurre la pareja. La última clase es la clase Lexicon que contiene una Hashtable con los nombres de entidades en inglés y otra con los españoles. También contiene una ArrayLyst con una entrada por cada artículo conteniendo una lista de NEs que aparecen en este artículo.



5.1 Diagrama de clases

El diagrama de clases de la figura 5.1 contiene los campos y operaciones más importantes. Las tres clases son serializables, que es muy importante para poder almacenar el lexicón en un archivo después de cerrar la aplicación. Este diseño permite alocar toda la información en la memoria dando rápido acceso de diferentes maneras. Por ejemplo se puede acceder a un nombre de entidad en concreto o se puede iterar sobre todos los NE que aparecen en un determinado articulo.

Para leer la información en memoria, la operación readNEs() lee la entrada de la fase de RNE. Luego la operación translate() ejecuta el algoritmo que genera el lexicón. Para cada nombre se ejecuta la función translate() y esta toma la decisión de cuál es la traducción del nombre.

Para decidir la traducción se toman en cuenta todas las traducciones candidatas del nombre. Se elige el candidata que coocurre el mayor número de veces. Si hay dos que coocurren el mismo número de veces no se puede hacer la traducción. Si el mejor candidata ocurre más veces que cierto umbral, del total de artículos en los que aparece el nombre, la pareja de nombres se incluye en el lexicón. Pero antes de incluirlo la misma comprobación se hace a la inversa para el nombre candidata. Esto es necesario porque puede que un nombre en inglés aparezca en cinco artículos y en

cada uno de ellos aparece una traducción candidata en español, pero si este nombre español aparece en otros 100 artículos aparte de estos 5 obviamente la traducción no sería correcta. Además si un nombre no aparece en un mínimo de artículos no se busca su traducción ya que la información se considera insuficiente.

La similitud entre dos nombres tiene más preferencia que el número de coocurrencias entre ellos. La primera razón para esto es que si los dos nombres son exactamente iguales entonces casi al 100% son una pareja traducida. No siempre, porque en algunos artículos en español, aunque haya una versión en español de cierto nombre se utiliza la versión inglesa. La segunda razón es porque aunque un nombre aparezca sólo en tres artículos y únicamente en un artículo comparable aparece una traducción candidata con un coeficiente de similitud muy alto es muy probable que estos dos sean una pareja válida. Incluso en estos casos se toma en cuenta el número de coocurrencias. Porque aunque la similitud sea muy buena si la traducción candidata solo aparece en 1% de los artículos lo más probable es que la traducción no sea válida. Así que existen dos umbrales para número de coocurrencias. Uno para parejas de baja similitud y otro para parejas de alta similitud.

Se hacen muchas pruebas para calibrar los umbrales. También se hace un análisis manual del data. Cambiando los umbrales se mejora la precisión pero se deteriora el coeficiente de recuperación o al revés. Los valores finales de los umbrales que se utilizan en la evaluación del lexicón son 0,7 para el coeficiente se similitud, como mínimo 5 ocurrencias para buscar una traducción y 51% de coocurrencias para las traducciones candidatas.

6 Evaluación del lexicón

Este capítulo da una evaluación numérica a los resultados obtenidos del sistema. Existen diferentes medidas para cuantificar los resultados y también diferentes partes del proyecto a evaluar. Las medidas más comunes utilizadas para evaluar proyectos en el área de recuperación de información son la precisión y el coeficiente de recuperación. La primera evalúa la exactitud y la segunda la completitud.

Dado que el proyecto tiene tres partes diferentes se evalúa cada una de ellas por separado.

6.1 Corpus comparable

Se pueden evaluar varios aspectos del corpus. Probablemente los más interesantes sean los relacionados al tamaño. El tamaño total del corpus es 65.8MB y contiene 27655 parejas de artículos con 4.998.462 palabras en inglés y 5.145.170 palabras en español.

Para otros corpus bilingües la precisión también puede ser una medida interesante pero en este caso la manera de emparejar los artículos es 100% precisa por lo tanto la precisión es de 100%.

También es importante medir la calidad del corpus para la tarea en cuestión. Cuantos más nombres contenga el corpus mejor es su calidad. Estos nombres tienen que ser los mismos en ambas versiones de un artículo. Para calcular la calidad del corpus se compara el número total de nombres de entidades que aparecen en ambos idiomas con el número de nombres que coocurren tanto en la versión española como en la versión inglesa del artículo. En 30 parejas de artículos elegidos aleatoriamente y explorados manualmente se encuentran 354 nombres, de los cuales 159 aparecen en ambas versiones del artículo. Esto es un 45%. Esto significa que el 45% del total de nombres que aparecen en el corpus podrían ser incluidos en el lexicón. Este porcentaje es muy bajo y el problema es que los 55% restantes son datos inutilizables que pueden impedir el correcto funcionamiento del sistema.

6.2 Extracción de nombres

La evaluación del sistema de RNE es una tarea que consume mucho tiempo. Por eso en la evaluación se utiliza una cantidad de datos muy pequeña. Se examinan manualmente 35 artículos en los cuales son encontrados 256 nombres en los artículos en inglés y 268 en español. Alchemy

descubre 240 de los cuales 193 correctos en ingles así que la precisión estimada es de 80,4% y el coeficiente de recuperación es de 72,8%. Para los artículos españoles se encuentran 235 nombres de los cuales 190 correctos, o sea que la precisión es de 80,8% y la recuperación de 70,9%.

6.3 Construcción del lexicón

Se utiliza una muestra de 1000 entradas para evaluar el lexicón. De estas entradas 939 son correctas que significa que la precisión es de 93,9%. El resultado es bastante bueno pero hay que tener en cuenta que 746 de las entradas se escriben exactamente igual en español e inglés. Esto significa que la precisión para parejas que no se escriben de la misma forma es de 76%.

Es muy difícil calcular el coeficiente de recuperación ya que el número total de nombres no es conocido. Sin embargo puede ser estimado. Se sabe que el sistema de RNE descubrió 29.395 nombres únicos en inglés y 27.474 en español. De los nombres en inglés el 80,4% son correctos que son 23.634 y estos representan el 72,8% del número total de nombres que da 32.464 nombres en inglés. Para el español 80,8% son 22.199 nombres reconocidos correctamente que representan el 70,9% del total, el total son 31.310.

Ahora dado que solo el 45% del número total de nombres es común en ambos idiomas, si *C* son los nombres comunes y *T* es el número total de nombres entonces *T=C/0,45.* El número total de nombres *T=E+S-C* donde *E* son los nombres en ingles y *S* los nombres en español. Entonces *C=E+S-T*, restando *T* esto es *C=E+S-C/0,45* o *C=0,45(E+S)/1,45=0,45(32.464+31.310)/1,45=19.791* nombres de entidades que pueden ser emparejados con su traducción.

Estos cálculos son muy poco exactos ya que los porcentajes de precisión y recuperación son obtenidos con muestras muy pequeñas.

El lexicón contiene 9.760 nombres de los cuales el 93,9% son correctos, o sea 9.165 que representa un 46,3% del total de 19.791 que es la recuperación del algoritmo. Este número puede parecer muy bajo pero en realidad no es tan malo ya que probablemente la mayoría de estos nombres sólo aparecen en un artículo, así que no sería posible traducirlos de forma automática.

7 Futuros trabajos

Se consideran muchas mejoras entre las cuales la utilización de programación con hilos para la recuperación de los artículos, ya que mientras se recuperan los datos de la Web se pueden hacer trabajos de procesamiento local.

Por otro lado una mejora en el diseño podría ser el uso de una base de datos en vez de tablas hash, para el almacenamiento de la información, en la construcción del lexicón. Serializar la tabla hash para hacerla persistente exige mucho tiempo que se podría ahorrar con el uso de bases de datos.

Para mejorar el resultado final y no sólo la velocidad del sistema sería necesario obtener más nombres en el lexicón y con más precisión. Para hacer esto se pueden mejorar por separado cada una de las diferentes partes del proyecto. Primero se puede buscar un corpus más grande y con más nombres y mejor calidad. Luego se podría probar el funcionamiento de alguno de los demás sistemas de RNE.

Por último se podría mejorar el alineamiento de nombres incluyendo nombres sinónimos en español e inglés. Por ejemplo Bush es sinónimo de George Bush y George W. Bush. En vez de crear un lexicón de parejas de nombres se puede hacer de parejas de conjuntos de nombres sinónimos. Esto puede ser una mejora considerable, ya que aunque US tiene que ser traducido como EEUU, Estados Unidos también es una traducción válida que tiene que ser utilizada en una búsqueda translingüe.

8 Conclusiones

La parte más importante es que se han cumplido todos los objetivos primarios. Se ha creado un lexicón bilingüe para traducción de nombres de inglés a español y viceversa, utilizando recursos Web. El lexicón puede ser actualizado a diario incluyendo los nuevos artículos de Euronews con una simple llamada a función.

Dado que los dos idiomas utilizan el alfabeto latino la clave para la construcción del lexicón fue utilizar la similitud de los nombres. Casi el 75% de los nombres no cambian de un idioma a otro y otro gran porcentaje se escriben de forma muy parecida. Para emparejar el resto de los nombres se utilizó el número de coocurrencias de las diferentes parejas de nombre inglés y español. Si la pareja ocurre muchas veces entonces es muy probable que sea una traducción válida. Esta técnica tiene varios parámetros que se fueron afinando con las pruebas y el análisis manual de los datos. Se obtuvo una precisión de 76% para los 25% de nombres que no se escriben igualmente en ambos idiomas.

La precisión final del lexicón es de 94% y la recuperación de 46% un resultado que se ve muy afectado por el sistema de reconocimiento de nombres.

Todavía hay mucho que desear ya que el lexicón no es muy grande y tampoco es muy preciso. Como se comenta en la sección de futuros trabajos sería interesante ver la posibilidad de tener más de una traducción para cada nombre.

9 Bibliography

1. *Message Understanding Conference - 6: A Brief History.* **Grishman, Ralph and Sundheim, B.** 1996.

2. **Kohen, Dr. Philipp.** *Statistical Machine Translation.* 2008.